AD 723214

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY *(Corporate author)*
Naval Ship Research and Development Center
Washington, D. C. 20034

2a. REPORT SECURITY CLASSIFICATION
UNCLASSIFIED

2b. GROUP

3. REPORT TITLE
AN EVALUATION OF ON-LINE INFORMATION RETRIEVAL SYSTEM TECHNIQUES

4. DESCRIPTIVE NOTES *(Type of report and inclusive dates)*
Final Report, December 1970

5. AUTHOR(S) *(First name, middle initial, last name)*
Theodore Wolfe

6. REPORT DATE
December 1970

7a. TOTAL NO. OF PAGES
60

7b. NO. OF REFS
12

8a. CONTRACT OR GRANT NO.

b. PROJECT NO. SR00308
Task SR0030801

c.
884-610

d.

9a. ORIGINATOR'S REPORT NUMBER(S)
3548

9b. OTHER REPORT NO(S) *(Any other numbers that may be assigned this report)*

10. DISTRIBUTION STATEMENT
This document has been approved for public release and sale; its distribution is unlimited.

11. SUPPLEMENTARY NOTES

12. SPONSORING MILITARY ACTIVITY
Naval Ship Systems Command
Code 0311
Washington, D. C. 20360

13. ABSTRACT

This report reflects current developments in on-line information retrieval systems and suggests goals for future developments of these systems. The review and comparison of the three major on-line information retrieval systems discussed in the paper reflect current achievements in working on-line information retrieval systems. A final section is devoted to suggestions for exploiting the potential of on-line remote access information retrieval and display systems.

14. KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
On-Line Systems						
Information Retrieval Systems						
Recon						
DDC Remote Access						
Tip						
Interactive On-Line Systems						
Man-Machine Systems						
Heristic Retrieval Techniques						
On-Line Display Systems						
Inferential Retrieval Techniques						
Terminal Input and Display						

DEPARTMENT OF THE NAVY
NAVAL SHIP RESEARCH AND DEVELOPMENT CENTER
WASHINGTON, D. C. 20034

AN EVALUATION OF ON-LINE INFORMATION RETRIEVAL SYSTEM TECHNIQUES

by

Theodore Wolfe

This document has been approved for public release and sale; its distribution is unlimited.

December 1970

Report 3548

TABLE OF CONTENTS

ABSTRACT

This report reflects current developments in on-line information retrieval systems and suggests goals for future developments of these systems. The review and comparison of the three major on-line information retrieval systems discussed in the paper reflect current achievements in working on-line information retrieval systems. A final section is devoted to suggestions for exploiting the potential of on-line remote access information retrieval and display systems.

ADMINISTRATIVE INFORMATION

This work was accomplished in the Information Retrieval Division of the Department of Apolied Mathematics under Project SR00308, Task No. SR0030801, Title: Remote Use of Computers for Naval Information Systems.

I. INTRODUCTION

This report presents a review and evaluation of three remote access on-line information retrieval systems and some ideas on what the capabilities of an ideal on-line information retrieval system should be. The three systems reviewed are the DDC Remote On-Line Retrieval System, the National Aeronautics and Space Administration RECON System, and the Technical Information Program (TIP) of the Massachusetts Institute of Technology.

The DDC system is examined in detail for two reasons. It is the newest of the three systems and reflects many of the latest advances in the utilization of on-line capabilities for information retrieval. It is currently in use at the Naval Ship R&D Center (NSRDC), and the author has had direct personal experience in its use and development.

Each of the three systems is reviewed on the basis of its operation during the last quarter of 1969. This paper does not reflect changes in the systems which have occurred since then.

II. THE DDC REMOTE ACCESS ON-LINE INFORMATION RETRIEVAL SYSTEM[1]

A. INTRODUCTION

The Defense Documentation Center (DDC) Remote On-Line Retrieval System is an experimental program which has the capability of querying the RD&T Work Unit Information System (WUIS) data bank from remote terminals. Its primary purposes are to provide direct access by the user to the DDC 1498 WUIS data bank, and to determine whether the relevancy or usefulness of stored information can be increased through a significant reduction in delivery time as compared to the batch mode of operation.

The experimental system consists of seven remote terminals which are used for testing and evaluating the concepts and methods applied. The seven terminals are deployed as follows:

Three at DDC (Cameron Station, Va.)
One at DDR&E (Pentagon)
One at Army (Arlington, Va. - Estimated operational date, July 1971)
One at Navy (Naval Ship Research and Development Center (NSRDC), Carderock, Md.)
One at Air Force (Andrews Air Force Base, Md.)
One at NSA (Ft. George G. Meade, Md.)

Each remote terminal is equipped with a Uniscope cathode ray tube (CRT), a remote buffer, and a remote printer which handle, on a time-sharing basis, all communications with the WUIS data bank. Circuits between DDC and remote terminals are secured through the use of Telecommunications Security (TSEC) equipment to protect transmission of classified material.

[1] References are listed on page 55

The DDC Remote On-Line Terminal equipment and Telecommunications Security equipment were installed at NSRDC in July, 1969. Instruction in the use of terminal equipment was provided by DDC and the terminal has been in operation ever since.

B. EQUIPMENT DESCRIPTION

1. Uniscope 300 Visual Communications Terminal

The Single Station Uniscope is contained in one case, and consists basically of a display screen, memory, control, input/output section, and character generator.

Display Screen. The display screen is the face of a cathode ray tube (CRT) with a viewing surface 10 inches wide and 5 inches high. A display format of 64 characters per line on 16 lines per display is provided, permitting a total of 1,024 characters to be displayed and/or to be held in memory. Spacing between characters is consistent from one end of the screen to the other, and the size and shape of the characters do not change in relation to their position on the screen. The character style maximizes legibility and readability. Each character is .150 x .113 inches and is readable from a distance of seven feet. Character brightness may be varied by the operator from 70% of brightness to full brightness. Since each character is repainted on the display surface 60 times each second, no flicker or jitter is perceptible to the viewer.

Operator Controls. The operator controls consist of an alphanumeric typewriter keyboard, cursor control keys, editing keys, indicators, function keys, and display controls.

Typewriter Keyboard. The typewriter keyboard very closely resembles the s.andard electric typewriter. Its purpose is to print messages destined for the computer. As a key is depressed, its character goes simultaneously to the display memory and to the display screen. It remains displayed until the operator edits and verifies the data before transmitting it to the computer.

Cursor. The cursor is a unique character that is displayed on the CRT at all times. The cursor indicates the location at which the next data character will be displayed It also indicates the starting position from which data will be transmitted to the computer Whenever the cursor is positioned on a displayable character, both will blink automatically. The blinking prevents the operator from losing track of the cursor when it is superimposed over a character. The cursor advances one space for each character that is typed and can be positioned by the cursor control keys. The cursor control keys are nondestructive and do not effect the information in display memory.

As the cursor moves to within eight positions of the end of any line, an audible alarm (which can be turned off) will momentarily sound and an indicator light which says "End of Line" will be lit The indicator will remain lit as long as the cursor occupies one of the last eight character positions. Additionally, as the cursor enters the bottom line of the display, the alarm will sound and the "Last Line" indicator will be illuminated.

Cursor Control Keys. The eight cursor control keys listed below are used when composing or editing displayed messages.

Scan Forward - This positioning key moves the cursor forward one space at a time, or ten spaces per second when held down.

Scan Backward - This positioning key moves the cursor backward one space at a time, or ten spaces per second if held down.

Scan Up - This positioning key moves the cursor up one line at a time, or ten lines per second if held down.

Scan Down - This positioning key moves the cursor down one line at a time, or ten lines per second if held down.

Note: The repetition rate of ten spaces per second is actually adjustable from five to twenty-five spaces per second and may be preset anywhere in this range.

Cursor to Home - The key repositions the cursor to the first character position on the display.

Return - The return key is similar to the carriage return on a typewriter, and positions the cursor to the first position of the next line.

Space - The space bar is in the position normally occupied by the space bar of a standard typewriter keyboard and moves the cursor forward one space for each depression.

Tab - This is a special cursor positioning key that moves the cursor forward until a special tab stop character is detected in the display memory. If a tab stop character is detected, the cursor will stop one character beyond it. If no tab character is found, the cursor will stop at the end of the display.

Back Space - The back space key is similar to the back space key on a standard typewriter and moves the cursor backwards one space for each depression.

Editing Keys. Five editing keys are used to correct or change data that has been input at the keyboard or received from the computer. The use of these keys is straightforward.

Character erase

Erase to end of line

Erase to end of display

Insert

Delete

Function Keys. Forty special function keys are available. Five are located above the numeric row of the typewriter keyboard The rest are located to the right of the typewriter keyboard. Only five of the special function keys are currently programmed for use within the WUIS Retrieval System.

Data Control Keys. There are two data control keys, the transmit key, and the message waiting key.

The transmit key causes data to be transferred to the computer. When this key is pressed, all data (64-character maximum, including spaces) on the same line and to the left of the cursor will be transmitted. The keyboard is locked out from further data entry until the message transmitted has been accepted by the computer.

The message waiting key is used in conjunction with unsolicited messages.

Indicator Lights. There are six indicators located above the keyboard. Their labels and functions are as follows:

Last Line - Lights when the cursor is positioned anywhere in the last line of the display.

End Line - Lights when the cursor is in any of the last eight character positions of any line.

Fault - Lights whenever a parity error is detected in the message that is being received from the computer.

Hi-Temp - Lights to warn the operator that the internal temperature is exceeding the normal limit.

Message Waiting - Lights whenever an unsolicited message is to be received from the computer.

Wait - Lights during the time a message is being transmitted or received.

Audible Alarm. An audible alarm will sound to alert the operator to three possible conditions. A single beep is sounded when the cursor moves into the 57th character position of any line and also when the cursor moves into the first character position of the last line. The alarm also sounds intermittently whenever an unsolicited computer message is waiting The alarm is turned off when the message waiting switch is depressed. It also may be adjusted so that it does not sound at any time.

Display Controls There are four display controls in the upper right hand portion of the keyboard These controls are described as follows:

On-Off - The on-off switch requires a passkey to operate It applies power to the Uniscope terminal and puts it in an operating state.

Focus - This control is used to focus the characters on the display screen

Louder - This control varies the volume of the audible alarm

Brighter - This control varies the brightness of the characters being displayed

Display Speed. The display generates five lines per second, or 320 characters per second at the rate of 3200 words per minute The transmission rate is dependent on the transmission rate of the modem equipment.

Display Memory. The display has a 1,024 character mem ry which transmits 256 characters at a time to the buffer for printing.

2. The Remote Buffer

The remote buffer acts as a control unit for the pagewriter. It provides the interface between the display scope and the pagewriter and holds display data, transmitted from the display for printing, in storage memory for the pagewriter. Data from the display scope memory is released to the buffer storage memory at the rate of 256 characters at a time

3. The Pagewriter

The operation of the pagewriter is similar to that of a teleprinter in that it produces hard-copy of incoming messages. The pagewriter is not program controlled and a release key on the display keyboard must be depressed to activate it for a hard-copy of the screen display. The pagewriter is mounted on a free standing pedestal which houses the pagewriter's power supply.

Printing Speed. The pagewriter operates at a speed of 1500 characters per minute with an average of 250 words per minute at the rate of 50 lines per minute.

Operator Controls and Indicators. Operator controls and indicators consist of six push-button switch/indicators and a mechanical adjustment control. Controls are described as follows:

Off Switch/Indicator - Depressing this switch indicator turns DC power off and lights the indi ator.

On Switch/Indicator - Depressing this switch turns DC power on and lights the indicator

Ready Switch/Indicator - Depressing this switch generates a general clear and ready signal to the remote buffer and lights the indicator

Error Indicator/Form Feed Switch - The indicator lights to indicate a data error, right margin error, or missing "remote" data carrier Depressing the switch feeds paper until the switch is released

Change Ribbon Indicator/Forms Out Indicator/Switch - The Change Ribbon indicator lights when the ribbon reversal counter equals zero, indicating a ribbon change condition. The Forms Out indicator lights when the unit is out of paper Depressing the switch when the indicator is lit clears the audible alarm

Text Switch/Indicator - Depressing this switch prints test characters (8 s) and lights the indicator

Print Phasing Thumbwheel - This adjustment positions ribbon for optimum printing legibility

C. SEARCH CAPABILITIES

1 Data Files

The WUIS data base consists of a Summary Data File (Direct File) and an Index File (Inverted File) which are monitored periodically by off-line file maintenance programs to insure current, accurate data and file compatibility

The Summary Data File (Direct File) is the master file and contains all the data fields pertinent to each WUIS summary. The file is organized by accession number; all the information for a given summary is located directly under the accession number of the summary. All Direct File data available to the user are described in the "RD&T Work Unit Information System Data Input Manual," DSAM 4185.5.

b. Index File (Inverted File/I.F.)

The Index File (Inverted File) is a retrieval file which consists of selected data fields extracted from the Summary Data File (Direct File). The data fields available for searching by means of the Index File are as follows:

1498 Field Nos.	Field Description	Inverted File Roles (DFID)
1	Agency Digraph	04
4	Kind of Summary	05
5	Security of Summary	58
10a	Primary Program Element No.	06
	Primary P.E. (1st 2 characters)	51
	Primary P.E. (1st 3 characters)	52
	Primary Project No.	07
	Primary Project & Task No.	08
	Primary Project, Task, & Work Unit No.	09
10b	First Contributing P.E.	10
	First Contributing P.E. (1st 2 characters)	53

	First Contributing P E (1st 3 characters)	54
	First Contributing Project No	11
	First Contributing Project & Task No	12
10c	Second Contributing P E	13
	Second Contributing P E (1st 2 characters)	55
	Second Contributing P E (1st 3 characters)	56
	Second Contributing Project No	14
	Second Contributing Project & Task No	15
12	Scientific/Technical Area (COSATI Code)	20
15	Funding Agencies (all)	21
16	Performance Method	23
17b	Contract/Grant Number	25
17c	Contract/Grant Type	24
17d	Contract/Grant Partial Code	57
17e	Contract/Grant Kind of Award	22
19c	Responsible Individual	28
19d	State/Foreign Country Code	27
	State plus Congressional District Codes	41
	DOD Organization Source Code - Six digit identification code of all WUIS source names	26
20c	Principal Investigator	31
20e	Principal Investigator Social Security No	16
20f	Associate Investigator (First)	32

20g	Associate Investigator (Second)	33
	State/Foreign Country Code	30
	State plus Congressional District Code	42
	Performing Organization Type Code	34
	Performing Organization Source Code	29
22	Keywords	
37	Descriptors - DDC assigned descriptive terms	
38	Identifiers - DDC assigned terms denoting specific equipments, projects, etc.	

The Inverted File is organized by type of data field and in accession number sequence within each data field.

b. Special File

The Special File is a temporary file created by the user for temporary storage of selected accession numbers resulting from searches. Its purpose is to allow the merger of the results of several separate search strategies so that their output may be combined into one result for batch processing. Duplicate answers which result from combined search results are deleted as part of batch processing.

2. Inverted File Search Capabilities

The Inverted File is made available for searching by means of Function Code "S". The following search qualification options are available for use on the Inverted File:

a. Hierarchy (Generic Help)

The hierarchy or generic help search qualification option is based on thesaurus structure. A thesaurus is primarily a dictionary of synonyms, and serves as a guide to the selection (and spelling) of a term which, among a group of synonymous terms, has been authorized as the descriptor for indexing the concept involved. In addition to guiding the user to the authorized descriptor among a group of synonyms, the thesaurus also lists under each authorized descriptor other authorized descriptors which are generically or closely related to it. The purpose of listing generic and related terms under each authorized descriptor is to guide the thesaurus user to the most specific authorized descriptor available for a given concept.

The designation of generic relationships is based on the hierarchical structuring of descriptors within families and indicates a broader or narrower relationship between descriptors. A broader term is a class term; for example, iron alloys is a broader term than steels. A narrower term is the reciprocal of a broader term and refers to a term that is a member of a class. For example, the terms 'steels', 'gray iron', and 'mottled iron', are narrower terms of the class or broader term, iron alloys. Iron alloys is, in turn, a member of the class. 'metals', and is listed as a narrower member or term of the class 'metals'.

This highly organized structuring of generic relationships between authorized thesaurus descriptors results in a powerful and efficient retrieval option for expanding the scope of a search.

Combining the symbol for the hierarchical search option with an authorized thesaurus descriptor will automatically expand the search to include not only all accessions posted to the descriptor but also all accessions posted to all descriptors designated by the thesaurus as generically narrower to the original descriptor.

Since the DDC Thesaurus is not available as an on-line file, the narrower terms themselves are not actually used to retrieve. Instead, a distinct descriptor is sought, i.e., $ plus descriptor. This is a separate descriptor distinguished from the normal descriptor by the preceding ($). ($) plus descriptor postings for broader class terms are automatically generated whenever a narrower class term is posted. Thus the combination of the hierarchical search option qualifier, ($) plus descriptor, is actually a request to retrieve all accessions posted to the term by indexers and all accessions posted automatically to the descriptor as a result of its narrower term posting.

b. Weighting Option

When a document is indexed, the descriptors representing the main subjects are indicated and differentiated from the set of associated descriptors by preceding the main subject descriptors with an asterisk (*) or weighting symbol.

Conversely, the search analyst may wish to increase the relevance of his answers by limiting his search to main subject entries by means of the weighting option. The weighting option is initiated by preceding a search term with an asterisk (*) weighting symbol and is effective only when used in conjunction with authorized thesaurus descriptors.

c. Masking Option

Weighting and hierarchical options are limited to Inverted File searches involving authorized descriptors. The masking or prefix option is not limited to the thesaurus vocabulary and allows the use of shortened terms. A search using the masking option will match the field value of its entry against every field value on the Inverted File except dollar amounts and dates. To avoid needless and time consuming searches the computer checks the length of the field values associated with a masking option operator. Any masked field value of four characters or less generates a scope display message cautioning the user and halts the search initiation until an override operator is submitted by the user.

d. Term Role (Field Code) Option

This option is used to search known data fields on the Inverted File, other than descriptors, identifiers, or keywords. The data fields on the Inverted File are identified by two-digit field codes (Data Field Identifiers) [illegible] e" codes. An Inverted File data field search is implemented by preceding the field value to be searched with a ? symbol (operator) and the appropriate two-digit "role" code (DFID).

e. Boolean Connectives

Once the search terms and the options or qualifiers to be used in conjunction with them have been selected, they are assembled into a search pattern or logical sequence by means of the normal Boolean connectives (AND, OR, NOT).

Boolean "OR". The "OR" connector is the most basic of Boolean connectors and is so frequently used that it is presumed to be present whenever two or more search terms are listed without an intervening Boolean connector. An "OR" logic pattern will retrieve the sum of all hits which satisfy any one of a set of alternative search terms. Duplicate hits are merged in the search result statistics.

Boolean "AND". The "AND" connector is used for coordinated search strategies and requires the satisfaciton of at least two separate conditions before an answer is accepted as a hit. The number of conditions which must be met is always one more than the number of "AND" connectors used.

Boolean "NOT". A Boolean "NOT" connector is used for exclusion. Answers which include any of the conditions listed under the "NOT" connector will be excluded from the final search result. The "NOT" connector is usually used as a final condition of an "OR", "AND", or "AND/OR" search logic pattern.

Combination Boolean "AND/OR". The use of an "AND" connector between sets of alternative "OR" search terms combines the advantages of both "AND" and "OR" logic. Its use permits considerable expansion of search coverage, particularly when hierarchical and/or masking qualifiers are used with the search terms involved.

3. Selected Direct File Search Capability

The search analyst may wish to further refine the results of an Inverted File search by qualifications on data fields not available for searching on the Inverted File. The "Q." function code opens the Direct File data base of documents selected by a prior Inverted File search and permits them to be searched by means of comparison values.

Direct File data fields may be qualified by the following comparison value symbols:

EQ = Equal

NE = Not equal

LE = Less than or equal

GE = Greater than or equal

GT = Greater than

The search analyst may further qualify prior Inverted File search results by comparison value searches on Direct File data fields such as dates or funding. The documents which do not meet the comparison values used are eliminated from the search. A final search statistic is displayed and the accessions which meet the qualification are available for further manipulation by other system capabilities.

Boolean connectors are available for use in conjunction with Direct File comparison values. Direct File comparison value searches must be formatted according to a format formula. Error messages are displayed if either a comparison value symbol or a Direct File data field identifier code is illegal.

D. TOTAL SYSTEM CAPABILITIES

In addition to its search capabilities the DDC Remote On-Line Retrieval System has many other system capabilities which can be called into use by their assigned function codes to act upon search results or to aid the user in formulating his search strategy or search display.

The total list of system capabilities may be divided into three types: action codes, auxiliary codes, and service codes.

1. Action Codes

These codes identify system capabilities which initiate and produce a user-sought response. The S., Q., and 7. function codes belong in this category as well as all action requests for search result display, sorting, and transfer to off-line printing and sorting.

Each file, Inverted, Direct, and Special, has its own set of display, sort, and batch transfer action codes. All three sets of file action codes which act upon search results have the same functions and capabilities, but are file dependent in their implementation.

Display Actions. Action codes which redisplay search questions and search statistics can be implemented only in conjunction with Inverted or Direct File searches. The Special File is a storage file of selected search results and does not need this type of redisplay. All other search result action codes may be implemented on any of the three files.

Common to all three files is the ability to display either search result accession numbers or the full display of each search result work unit. Search result work unit displays may be formatted according to standard display patterns or tailored to display any field in any format desired. Answers may be viewed on the scope either in a continuous mode with no pauses between items displayed or in a non-continuous single frame mode with page changes on request. Accession sequence may be requested in either ascending or descending order. Forward or backward browsing of work unit displays is available in a non-continuous mode and includes a skip value option which may be set from a value of 2 to 9999 as determined by the user. Display formats and browsing modes may be changed at will between item displays.

Sorting Actions. Sorting Actions are limited and dependent on whether the items to be sorted are for scope display and terminal printing, or for off-line batch processing. A maximum of three sort fields may be specified for terminal display or printing. The maximum sort for off-line batch processing is four sort fields.

Off-Line Batch Printing of Search Results. Eighteen printing formats are available for off-line batch printing requests from terminal users. A maximum of six printouts per search result is available. The kinds of printouts may be any combination of copies per format or number of formats not exceeding a total of six. Sorting of off-line batch printing is limited to four fields. Accession number sequence is always in descending order.

Display of a Known Accession Number. If an accession number is known, its summary may be called for display by use of the unique action code (W.). Action code (W.) acts on a single accession number. It can be used to reference the summary of a known work unit or to view an accession listed as one of the results of a search request. Accessions called into display by action code (W.) may be separately transferred to batch off-line processing of single work units in multiple formats. If the user has a list of accession numbers, each number may be requested individually by means of the (W.) code and transferred, individually, to the Special File where the accession numbers may be accumulated and transferred to batch processing as a package.

2. Auxiliary Codes

The three auxiliary codes, P, Y, and N. are used to respond to scope display messages concerning display or search continuation.

3. Service Codes

Service codes are used to call reference data displays such as display codes and formats and data field code identifications, and to reference examples of how to formulate search, sort, or batch requests.

In summation, the DDC Remote On-Line Retrieval System Capabilities cover every operation needed for information retrieval and display. A complete list of systems capabilities follows.

Action Codes

S.---------- Implement I.F. Search

SE.------------- Redisplay I.F. Search Statistics

SO.------------- Redisplay I.F. Search Question

A.-------------- Display Accession Number List From I.F. Search

AD.------------- Display WUIS Items From I.F. Search

AB.------------- Transfer I.F. Search Results To Batch System

AST.------------ Sort Inverted File Search Results

Q.---------- Implement D.F. QUAL Search Using I.F. Search Results

QE.------------- Redisplav D.F. Qualification Search Statistics

QO.------------- Redisplay D.F. Qualification Search Question

X.-------------- Display Accession Number List From D.F. Qual Search

XD.------------- Display WUIS Items From D.F. Qualification Search

XB.------------- Transfer D.F. Qual Search Results To Batch System

XST.------------ Sort Direct File Qualification Results

TZ -------- Build Special Accession Number File

Z.------------- Display Accession Number List From Special File

DZ ------------ Display WUIS Items From Special File

ZB ------------ Transfer Special File Results To Batch System

ZST ----------- Sort Special File Items

W.--------- Display Single Known WUIS Item

<u>Auxiliary Codes</u>

P.--------- For Paging Screen Displays

Y --------- For A Yes Response Or To Ignore The Message And Proceed With The Function

N --------- For A No Response

<u>Service Codes</u>

FD -------- Request CRT Item Format Change/Continuous Display Mode

LL -------- Display Site Activity Log

SS.-------- Display Sample I.F. Search Pattern

ST -------- Display Sample Sort Request

SQ -------- Display Sample Qualification Pattern

SB -------- Display Sample Batch Interface Request

I --------- Display D.F Fields/I.F. Role Numbers

U --------- Display CRT Item Formats

C.--------- Display of Total System Capabilities

OPT.------- User Guide To Sequential System Processing

E. CRITIQUE

The DDC Remote On-Line Retrieval System has been very successful. It allows the user to interact with the data base and modify his query based on feedback. It does this quickly, easily, and relatively cheaply in terms of computer time costs. With this new system DDC has overcome the basic weaknesses of batch processing No longer is there a need for intermediaries who may distort or misinterpret the user's intention The rigidities inherent in batch processing, which make query modification difficult, have been eliminated. The time saved is immense, and user confidence in the relevance of his answer is much improved.[2]

Despite well merited praise, there are flaws in the present DDC On-Line System. Two of them are major: lack of an on-line thesaurus file, and lack of system reliability Of these two, the lack of an on-line thesaurus file is the most hampering to the user Despite printed reference aids, the user must spend a great deal of time selecting appropriate descriptors to use in his search Without a thesaurus, his effort to discover the right synonym and available narrower terms is a hit or miss process This process of discovery can be time-consuming, exasperating, and misleading There is a very real need for an on-line thesaurus file to guide the user to the actual terms which have been used to index his areas of interest. DDC, by its omission of access to an on-line thesaurus file, has seriously curtailed the system's utility for subject term searches

System reliability, the other major problem, is a constant one with any new computer system, particularly one which has been subject to as much revision as has the DDC Remote On-Line Retrieval System. Even so, down-time shou not be as much of a daily hazard in the life of the user as it has been and continues to be with respect to both hardware and system operations.

Another area which should be improved is off-line batch request capabilities. At present this area of system capabilities is rigid and limited. Although 18 print formats are available, seldom do any of the 18 formats, other than the 1498M format, meet the listing requirements of a NARDIS request. A free form report generator capability similar to that available for special on-line display formats should be made part of the batch request capability. Sorting ability, which is now limited to four fields for batch processing and three fields for on-line displays, should be expanded to a more optimal number of fields for either capability. And, finally, the report generator should have a summing capability for at least three fields in order to generate summary report requests for on-line display or batch printing.

III. THE NATIONAL AERONAUTICS AND SPACE ADMINISTRATION RECON SYSTEM[3]

A. INTRODUCTION

RECON, an acronym which stands for REmote CONsole, is the name of the real-time, on-line, time-shared, information retrieval system used by the National Aeronautics and Space Administration. RECON was started in February, 1969, and now has 21 terminals installed at various NASA centers throughout the United States. All RECON terminals are linked by leased telephone lines to an IBM 360, Model 50, computer at the NASA Scientific and Technical Information Facility in College Park, Maryland.

B. EQUIPMENT DESCRIPTION

Each RECON terminal consists of the following pieces of equipment:

1. CRT Display

The present RECON CRT is a portable model with an 8 x 12 inch display screen.

2. Keyboard

The keyboard, also portable, is connected to the CRT display by electric lead wires. The separation of the keyboard from the CRT display allows for easy positioning of either unit to suit the user's needs. The keyboard consists of standard electric typewriter keys, plus special function keys, and cursor control and editing keys.

3. Printer

The RECON printer is a standard Western Union teletype machine. Its printing speed is 15 characters per second or 900 characters per minute. It operates in a receive only mode and prints on operator or remote computer command.

4. Control Unit

The control unit contains a data modem and a CRT printer buffer unit. The data modem controls signal transmissions to and from the remote computer. The CRT printer buffer unit controls signal transmission between the CRT display and the printer.

5. Response Time

With all twenty-one terminals in simultaneous use, message response time to and from the computer is usually ten to twelve seconds or less. While RECON may claim priority, if necessary, for computer access, it usually operates in a multiprogram environment which permits computer access to other programming requests.

C. SEARCH CAPABILITIES

The main data base available to RECON users is a bibliographic citation data bank describing the more than 600,000 documents housed at the NASA Scientific and Technical Facility at College Park. Each document citation includes catalog data, such as author, title, report number, date, publisher, and project number, i.e., the standard reference data used in bibliographic citations. Each document citation also includes the keywords assigned by the author of the document, the index terms assigned by the NASA indexing staff, and special identifiers, assigned by the author or indexer. Textual data, such as abstracts or report summaries similar to those available on the DDC On-Line System, are not included as part of NASA bibliographic citation data base. The purpose of the report citation file is to retrieve bibliographic citations to literature which may be obtained in microfiche form in the libraries and offices where RECON terminals have been installed. Since the reports themselves are easily available, summaries and abstracts are not needed.

1. Data Base

The data base comprises a linear file, an inverted file, and a thesaurus file. Formalized catalog data entries such as author, dates, publishers, project numbers, etc., are assigned identifier codes on the inverted and linear files and may be searched in much the same fashion as similar catalog data is searched on the DDC Inverted and Direct Files.

2. Subject Searching

The main difference between the search capabilities available to DDC On-Line users and RECON users is in their subject search capability. The DDC vocabulary available for subject searches is based on three sources: DDC thesaurus authorized descriptors assigned by the DDC staff; keywords assigned by resume authors; and special identifiers assigned by either DDC indexers or resume authors. The same three types of postings are made for the bibliographic citations on the NASA file. The difference between the two term files is that only part of the DDC term file is selected from a controlled thesaurus vocabulary, whereas all term postings (keywords, identifiers, and index terms) on the NASA file are selected from the controlled vocabulary of the NASA thesaurus.

It is this factor, the thesaurus controlled vocabulary, that distinguishes the RECON system and gives its user a great advantage over the DDC On-Line user. The DDC On-Line user may employ the hierarchial option only for authorized thesaurus descriptors. Appropriate items posted to unauthorized descriptors are not available for hierarchial retrieval. In contrast, because of its controlled vocabulary, all descriptors on the RECON data file are available for hierarchial searching regardless of their initial assignment source, i.e., original author or NASA indexer. As a result, a RECON user may expand his search to include hierarchially narrower terms and is assured of complete coverage of any type of descriptor which may have been assigned to any citation.

3. The On-Line Thesaurus and Usage File

Another major feature of the RECON system is its on-line thesaurus and usage file. Unlike the DDC On-Line user, the RECON user does not have to locate the appropriate subject terms for his search on a hit-or-miss basis. The RECON user merely enters the term he thinks is appropriate. His term is matched against the thesaurus and the corresponding authorized thesaurus term is displayed along with a count of the number of citations which have been posted to it. To determine the narrower terms available for a previously selected term, the RECON user enters an expand command, and the narrower terms and their usage counts are displayed below the original term and its usage count. The user may then either discount the narrower terms as part of his search formulation or indicate his wish for their inclusion in his search.

This elegant and efficient location and display of appropriate search terms, their usage, and available hierarchical search expansion capabilities is a major milestone in the development of on-line information retrieval systems. It represents a great advantage available to any system which uses a well-developed, thesaurus-controlled vocabulary for indexing and retrieval.

4. <u>Boolean Connectives and Query Language</u>

The Boolean connectives used in RECON are not words, but a combination of function keys and a mathematical formula. Symbols are used to represent the Boolean connectives between sets of terms and numbers are used to indicate subsets of terms. The answer response is a final subset number which includes only those items which meet the parameters of the formula. This method of indicating search levels and Boolean connectives may be efficient, but it is not clear or easy to use without added instruction.

The lack of language as a means of communication with the computer is most conspicuous in query formulation, and is a major fault throughout the RECON system Despite the efficiency and elegance of its retrieval capabilities, RECON's use of function keys and numbers for dialog between the user and computer reduces its communication ability. DDC's on-line ability to recognize words in place of function keys is not highly developed, but compared to RECON's lack of language, DDC's query language represents a great advance in on-line user communication language and makes the DDC on-line system much easier to use.

5. Display, Off-Line Batch, and Sort Capabilities

Since its data base does not include textual material, RECON's ancillary display, sort, and batch capabilities are not as varied as those available to the DDC on-line system. It does, however, have the same capabilities as the DDC on-line system for display, sort, and batch capabilities that do not involve textual data. RECON also has a special file, as does the DDC on-line system, for selected answer retention and manipulation.

D. CRITIQUE

NASA's RECON system has been in successful operation since 1969. Originally limited to service in the Washington, D.C. area, it now has nation-wide coverage and gives users who are thousands of miles apart simultaneous, equal, efficient, and prompt access to the NASA bibliographic citation file. It was one of the earliest on-line information retrieval systems developed and quickly proved the superiority of on-line retrieval over query batch processing procedures.

Although the DDC On-Line System and the NASA RECON System have many similar search and display capabilities, their data bases are very dissimilar and the two systems should not be considered as comparable to each other in their retrieval capabilities.

The RECON data base, though very large, is relatively simply structured in comparison to the DDC WUIS data base. Each WUIS accession requires approximately three times as many data field entries as a RECON bibliographic accession. The RECON file is cumulative; once a bibliographic citation is entered on the file, no further data manipulation is required. The DDC WUIS file is also cumulative, but each accession and its accompanying hundred or more data fields is subject to constant revision and replacement. The search capabilities for either data base are similar, but the amount of retrievable information for each accession is much greater in the DDC System than the NASA System. On the other hand, the NASA data base contains a vastly greater number of searchable accessions than the DDC data base. Both retrieval systems represent milestone achievements in on-line information retrieval development.

IV. THE TECHNICAL INFORMATION PROGRAM (TIP) OF THE LIBRARIES OF THE MASSACHUSETTS INSTITUTE OF TECHNOLOGY[4]

A. INTRODUCTION

Because of its prestigious academic setting and extensive review and discussion in professional journal literature, the Technical Information Program (TIP) of the Libraries of the Massachusetts Institute of Technology is probably the most widely known remote access on-line information retrieval system. Developed in 1962 as part of MIT's Project MAC (Multiple Access, Time-Shared Computers), TIP's original purpose was to provide an on-line search and retrieval capability for bibliographic data from physics journals. Since then the Technical Information Program has been generalized to cover multipurpose applications of its abilities to other types of data files. The Technical Information Program continues its original purpose and at present its data base includes bibliographic data covering more than 30,000 articles from 38 physics journals.

The journal holdings cited start with the first issue of 1965 for each journal and the holdings are updated weekly. The file contains the following information for each article:

i - identification (journal, volume, page)

t - title

a - author(s)

l - author's institutional connection

c - citations and bibliography

The literature is arranged by journal and volume. The search language all the location and retrieval of a set of article citations that contain a given item or satisfy a set of conditions.

MIT's TIP differs from both DDC's Remote Access On-Line System and NASA's RECON System in that it does not employ a CRT display TIP also has retrieval capabilities not included in the two previously discussed systems.

B. EQUIPMENT DESCRIPTION

The Technical Information Program is one among a library of programs available to the approximately 30 users of the Compatible Time-Sharing System (CTSS) on the IBM 7090 computer used in M.I T 's Project MAC. Access to the computer is by telephone lines which link the IBM Selectric 2741 Consoles to the computer.

An IBM Selectric 2741 Console is a free standing unit with three parts.

1. Typewriter

The typewriter is similar to the standard IBM Selectric "golf-ball" typewriter. In fact, when the keyboard is not in use as a remote console, the typewriter may be used as a regular typewriter.

2. Control Unit

The control unit is located behind the typewriter and contains Dataphone switching equipment and some computer circuits The control unit also contains a switch which enables the user to use the typewriter either as a remote console keyboard or as a regular typewriter.

3. Telephone

A telephone line to the MAC computer is obtained by dialing into the system by standard pushbutton telephone located next to the typewriter keyboard. Access to a line is determined by the number of concurrent users of the system. The maximum limit of concurrent users is 30; priority is assigned in advance.

C. LANGUAGE

A distinctive and laudable feature of TIP is its use of English words instead of role or code numbers or symbols for data field identification. For example, to find the word "laser" within a title, the user types the command:

Find title laser

While this is far from being a natural form of sentence structure, it is a great deal easier to remember and formulate than having to refer to a location code for title data field identification or a format formula for field code and search value entry.

Users familiar with the TIP data file may use abbreviations for commands, fields, and journal titles. Commands and fields may be abbreviated to their first letter. Journal titles may be abbreviated to standard citation forms, standardized six-letter acronyms, or numeric codes.

D. SPECIAL FILE CAPABILITY

Another TIP feature is its OUTPUT SAVE capability. Each TIP user is allotted a portion of disk track storage space which he may use to save the results of a search for future referenee and further searching. A saved file may be searched in the same manner as the TIP library file. A TIP library user may make a comprehensive search of the TIP library for a broad subject area, save the search results in his own private file, and at his convenience make refined searches on various aspects of the subject at will. He can also save the subsearches as separate subject files.

Saved files are not stored as part of the TIP library program and can be manipulated by executive commands outside of TIP Having a subset of the TIP library as his own private file gives the TIP user facilities for information retrieval, storage, and manipulation to be envied.

E. DISPLAY AND OFF-LINE PRINTING

Any arrangement of tha five fields available may be designated by the user for on- or off-line printing. Sorting is automatically alphabetic according to journal title, author, article title, and numeric according to journal volume and page Off-line printing is done outside the TIP program and requires a store file command and a print request.

F. SEARCH CAPABILITIES AND CRITIQUE

The data base of TIP is very similar in structure and form to that of RECON. Both data bases contain bibliographic data without text. Both have similar search capabilities for Boolean coupling of search terms and data fields in various combinations. TIP, unlike RECON or DDC's On-Line System, does not rely on human processing for the assignment of descriptors, keywords, or special identifiers. TIP does not have a thesaurus or list of descriptors. Instead of terms assigned by indexers, TIP relies on the words included in artic' itles as a source of index terms. It can do this because the article titles are taken from journals which require their authoi to use titles which fully reflect article content. Since it cannot rely on an authorized list of descriptors for spelling of prefixes, suffixes, or phrase arrangement, TIP supplies its users with six options which they may employ to define the object of a search or a condition to be satisfied.

The types of search options are as follows:

. Prefix Match - Everything beginning with the desired prefix will be retrieved.

. Exact Match - Only terms which exactly match the desired term will be retrieved.

. Exact Suffix Match - Only terms which end with the suffix cited will be retrieved.

Masked Suffix Match - By indicating the number of participial, adjectival, or plural ending characters, a suffix search may be expanded to include many variations of spelling and meaning.

Words in Any Combination - All titles containing the terms cited, regardless of the order of their appearance within a title, will be retrieved

Words in Exact Order - Only titles which contain the terms cited in the exact order cited will be retrieved.

This array of term search capabilities gives the TIP system a very comprehensive and ingenious term retrieval system. In effect, it offers its user a word root stem and text search capability without encumbering the system with the elaborate and large word stem dictionary and comparison tables usually required to achieve either of these rare search capabilities.

To offset its lack of hierarchical search expansion capability, TIP offers its user a unique capability made possible by the formalities of its data base and data base source. Although TIP's data base does not include index terms based on a hierarchically structured thesaurus, it does include the references cited in each article's bibliography It is assumed that articles on the same subject will have bibliographic references in common even though their titles have no words in common. Therefore, to obtain a search depth and expansion equivalent to hierarchial search expansion, TIP allows its user to add to his search response all articles whose references include the same references as the article or articles obtained by his title term search. Expanding literature search depth by checking reference commonality is a venerable information retrieval technique, heavily used in manual literature searches. It is seldom, however, employed in computer-oriented information systems other than in the field of jurisprudence, in which historical relevance and precedence are primary concerns.

By relying on the formalities inherent in its data base, TIP has achieved a comprehensive, efficient, inexpensive, and easy to use information retrieval system. Unfortunately, TIP will work only for information systems which share its data base input reliability. The context of TIP is physics, a field of information which has a narrow and specific vocabulary. In addition, the physics profession closely governs its literature formalities. All the journals included in the TIP data bank require very similar reference citation formats and title reliability. Few professions have achieved the same commonality of form, format, and vocabulary as physics has in its journal literature. The rigidity of its data base is what enables TIP to achieve its ingenious search capabilities. It is this same requirement for data base rigidity that has prevented the universal adoption of TIP for computerized information retrieval in other fields of professional literature.

V. SUGGESTIONS FOR EXPLOITING THE POTENTIAL OF ON-LINE REMOTE ACCESS INFORMATION RETRIEVAL AND DISPLAY SYSTEMS

A. INTRODUCTION

The ideal remote access on-line information retrieval and display system would optimize the feedback capabilities of user-computer interactions to achieve the fullest possible use of the computer as an information retrieval and display tool. Current on-line information retrieval and display systems are mainly on-line program adaptations of batch processing techniques and do not exploit the inherent possibilities of on-line user-computer interaction, unhampered by the rigidities and time lag of batch processing techniques.[5]

The following paragraphs describe some of the main characteristics and capabilities which might be included in future on-line information retrieval and display systems.

B. BASIC TUTORIAL DISPLAY SEQUENCE[5,6]

The tutorial sequence should provide enough background and instruction to train a user completely unfamiliar with the system. It should explain the origin and content of the data base, provide instruction in query formulation, and thoroughly describe the system's capabilities, language, and limitations. In addition to reference displays of sample queries, data field identification tables, and display format models, the tutorial sequence should include all the material now provided in the form of printed instruction manuals. For the new user it should provide a computer-aided instructional course in query formulation using a sample data base to test the new user's skill and understanding of the system before allowing him access to the full data base. The importance of the tutorial sequence cannot be overstressed. It provides the basis for user-computer interaction and by doing so determines in great part the success of potential use of the system. A good tutorial sequence will create user self-confidence, increase the efficiency with which the system is used, and greatly expand the system s marketability.

C. SYSTEM CONTROL LANGUAGE AND PATTERN RECOGNITION CAPABILITIES

Ideally the system user would be unaware that the language he uses to communicate with the computer presents any problems at all. In fact, the user woul: probably prefer oral communication as the medium of interaction, and much work is being done at NSRDC .o develop this capability.* The user has no desire to learn a new language in order to communicate his wishes to the computer. His acceptance of the system is in large part determined by how little he must adjust his normal means of communication in order to use the system. The more natural the language, the greater will be the use and acceptance of the system. Ideally, the user should be able to address an on-line information retrieval system in the same way he addresses a librarian -- by a naturally formed

*The Speech Recognition Group in the Computation and Mathematics Department of NSRDC is headed by Dr. S. Berkowitz.

question or statement, such as, "What do you have on Henry VIII?", or "How much money is being spent on laser research?". The computer would then analyze the request, formulate and perform the search, and present the answer. This sounds like crystal-gazing, yet with on-line capabilities, it is actually possible to a surprising degree, even at the present stage of computer pattern recognition capability.

The system control language and pattern recognition capabilities employed in a user-computer interaction situation should have the following characteristics:

- The system should provide automatic search and display formulation capabilities based on pattern recognition, utilizing interactive user interrogation to determine automatic search and display pattern expansion or limitations. For example, if a user submits his question in an unacceptable format, the basic tutorial sequence would automatically be brought in for appropriate instruction. The instruction would then lead the user to automatic query analysis search and display programs. The automatic query analysis programs would include a program of computer interrogations to elicit from the user the parameters which he requires but omitted from his initial request because he did not understand the best manner of query formulation.

- For users who wish to bypass the automatic query and display programs, the system should provide two alternate vocabularies for system commands and data field identifications.

The basic vocabulary should be English words which reflect command actions or describe the data fields covered. A recognizably abbreviated version of the basic vocabulary should be made available as an alternate system control language for heavy users. The system should be capable of accepting commands and data identifications which combine both vocabulary forms.

- All query formulations should be automatically analyzed for scope and logic. If the analysis determines that the query formulation is questionable, it will automatically call in the appropriate level of user-interview and user-instruction programs. Ambiguities or questionable strategies would be clarified by interactive explanatory interview sequences. On resolution, the search would continue in its normal sequence.

D. SEARCH CAPABILITIES[1,3,4,5,6,7]

Search capabilities should include the standard options used in batch processing, such as:

1. Exact Match - Only items which exactly match the desired item will be retrieved.

2. Prefix Match - Every item beginning with the desired prefix will be retrieved.

3. Suffix Match - Every item ending with the desired suffix will be retrieved.

4. Combined Prefix and Suffix Match with Masking as Indicated - A combination of prefix and suffix and masking options allows the user to ask for middle of the term comparison.

For example, a search on the term, 'concept', using prefix, suffix, anu character masking indicators in combination, will retrieve the following terms:

concept
concepts
conceptualization
conception
preconcept
preconception
preconcepts
preconceptualization
misconcept
misconception

and other possible prefix and suffix combinations involving the basic word, 'concept'. In effect, the combination of prefix, suffix, and character masking indicators offers much the same retrieval capability as word root stem search systems which rely on dictionary files, and extensive prefix and suffix ending comparison tables.

5. Hierarchical Expansion - If desired, the search term can be compared against an on-line thesaurus file display and the search can be expanded to include synonymous, narrower, and/or related terms cited as appropriate by the thesaurus file. The thesaurus display can include a usage count for each of the appropriate terms shown. Hierarchical expansion capability is not limited to words. It can also be applied to numeric or alphanumeric classification systems which are hierarchically structured, such as a hierarchical decimal classification system.

6. Quality Comparison on All Range Variations

7. Boolean Connectives - AND, OR, and NOT connectives can be used as search parameters.

8. Limited Text Search - Any retrieval system with the foregoing search capabilities is also capable of limited text searching, provided the text data fields to be searched are arranged in a sequential format, so that the text may be searched on a word for word basis. The value of a limited text search capability depends upon the data base. Data bases which have one- or two-line text data such as titles can use text search as a fundamental retrieval capability. Data bases which contain a great deal of textual data may find text searching too time-consuming to merit incorporation as a search capability unless confined to searching titles or similar selected textual data areas.

In either case, text retrieval requires a combination of matching capabilities, Boolean connectives, and linkage indicators. Linkage is required in order to indicate word position acceptability. For instance, a text search on the phrase 'thin films' requires parameters to indicate which or how many of the following three word arrangements was acceptable as an answer;

thin films
thin magnetic films
films of thin people

Obviously, the merit of limited text searching is dependent on the deductive reasoning and linguistic prowess of the search analyst. Backed by an adequate on-line thesaurus, however, and hierarchical expansion of term input, limited text retrieval can be developed into a powerful ancillary search tool for use with one- or two-line text data bases such as bibliographic data banks.

9. Weighting - Weighting retrieval capabilities are currently dependent on values assigned by indexers as part of data input. Any data base which includes weighting as an input value can incorporate weighting as a search capability. Weighting, and its allied value, linkage, need not be dependent on man-assigned values but can be determined by computer-performed statistical inference, an on-line capability which is discussed in the next section.

E. AUTOMATIC QUERY FORMULATION BY MEANS OF STATISTICAL INFERENCE ANALYSIS AND HEURISTIC OPTIMIZATION [5,6,8,9,10,11,12]

Statistical inference is an on-line capability particularly appropriate to information retrieval. Utilizing statistical inference programs which can analyze in advance the hit probabilities of every possible search strategy, the user could cite known answers for computer analysis and ask the program to determine the best search strategy to retrieve documents which contain similar data. The degree of similarity required could be specified by the user on a percentage basis, i.e., 90%, 80%. The procedure used for statistical contingency analysis* and percentage matching could also be adapted to analyze and assign percentage weighting of input data. The actual analysis could be based on the number of hits which a full match of every data field would obtain. A succession of matches, subtracting one data field at a time, would heuristically derive the optimum search strategy.

To increase efficiency, automatic search formulation analysis could be programmed to ignore various data fields such as accession numbers to avoid pointless comparison. A standard group of comparison fields could be determined in advance with additional fields added as search parameters by user feedback. User feedback acquired by computer interrogation of the user would be utilized to weight search parameters, such as main subject concept and logistic data requirements which must be met in order to satisfy the search request.

If two or more known answers are cited by the user, the automatic search analysis program would compare them for data field commonality, weight the search by means of user-identified essentials, and proceed from there to determine the optimum search strategy. The same search formulation analysis program combined with user interrogation programs could be used to initiate search strategies without using known answers as a basis for search formulation.

*Statistical contingency analysis refers to the statistical frequency analysis of the occurence of items within a given context or environment.

The key factors in automatic query formulation by means of statistical contingency analysis and heuristic programming techniques are on-line user feedback capabilities and the incredibly vast and almost instantaneous computational capacities of the computer. Optimum search strategy formulation by computer is not achieved by deduction, but by trial and error, or heuristics, a very inadequate procedure for humans. When done by computer, trial and error computation is done so massively and instantaneously that an optimum search strategy can be heuristically obtained by computer in less time than it takes for a human analyst to enter his own search strategy into the on-line retrieval system. Automatic search formulation based on on-line user feedback, statistical contingency analysis, and heuristic optimization procedures will undoubtedly develop into the primary search formulation method for all future on-line information systems regardless of their data base content. The concept involved is extraordinarily practical in terms of potential use and personnel cost reduction. Eventually, a generalized query formulation program will be developed and made adaptable to every kind of data base. Query coordinators and search analysts may prove redundant.

F. DISPLAY CAPABILITIES

Ideally, display capabilities should be the same for all output media and should be able to meet every foreseeable user requirement. To accomplish this goal, the ideal system would utilize a universal report generator capable of displaying, sorting, formating, and summing as required any data supplied to it, regardless of the data source.[5] In addition to being capable of accepting any design requirement, the report generator should be able to supply standard displays and report formats pertinent to search formulation, search result statistics, and search review or document browsing.

Search result statistics displays should be in the form of a matrix showing hits per search parameter and should be ranked according to number of parameters met. In conjunction with this display the user should be able to request accession displays for each rank of parameters met by specifying which rank of answer sets he wishes to see.

On-line print requests should be capable of by-passing the CRT display prior to printing. The user should be able to hold a CRT display sequence at will and should also be able to have any portion of a CRT display printed at will. He should be able to alter display data prior to a print request and have the display printed as altered.

Print or display requests involving search result displays should be capable of accepting user specified headings and data field labels in place of standard headings or labels. User designed formats should automatically display a sample format based on the user's design and should be capable of accepting format design revisions by the user.

The universal report generator should also be able to supply standard graphic displays of statistical data. Graphics capabilities should include color choice; graph superimposition; and a choice of graph type, e.g., axial, bar, or pie chart.

G. SPECIAL FILES

1. Temporary Storage File - A temporary storage file should be available for selected answer transfer and storage. Selected answers stored on the temporary storage file should have the same search and display capabilities as the main data base.

2. Query Library File - Permanent storage space should be allotted for a Query Library File. The Query Library File should be searchable and capable of displaying its data of query formulations, search statistics, and report generator formats.[5] Appropriate data would be automatically generated and transferred to the Query Library File whenever a query is addressed to the main data base. The Query Library File would be used to maintain statistics on system use and to store formulations and formats

..ed for recurrent queries such as standard recurrent subject bibliographies or user interest profiles. Whenever its allotted space nears capacity, the Query Library File should automatically submit a self-printout for review and selective data re-storage.

H. EQUIPMENT REQUIREMENTS

The type and extent of remote terminal equipment depends upon the user's needs. Casual use would not require a CRT or high speed printer and, in many cases, the only required equipment would be a typewriter console and data modem.

Heavy and complex usage would require the following equipment triad:

CRT Display

Typewriter Keyboard and auxiliary input devices

High Speed Printer

1. CRT Display - The ideal CRT display would include a full page of text, displayed clearly and legibly, and requiring no eye strain. Display speed should be user governable to accommodate various reading speeds and data comprehension rates. All display controls should be located on the front of the display and be easily accessible. The display equipment should include a projection device for wall screen enlargement of CRT displays for large audience viewing. The CRT unit should be as compact and light as possible and easily moved.

2. Keyboard and Display Control - The keyboard should contain the standard electric keyboard with cursor and display control keys located on the user's right for easy access and use. If the terminal is to be used for remote batch input, the keyboard should accept auxiliary input devices such as magnetic disk or cassette tapes, punched cards, data phones, etc.

3. Printer - The printer should be a silent, high-speed, non-impact device capable of producing alphanumeric or graphical printouts in single or multiple copies as needed.

I. SUMMARY

On-line interaction between the user and the computer opens the door for the use of the immense computational capabilities of the computer as a tool for information retrieval methods based on statistical analysis of user feedback. User-computer on-line interaction also offers possibilities for display editing and data manipulation far beyond current program limitations of on-line display and report generators. Specifically, the ideal on-line information retrieval and display system should include all the capabilities of batch processing enhanced by the directive and interaction capabilities of on-line systems for user-computer information exchange.

ACKNOWLEDGMENTS

Many people deserve credit for their guidance and assistance in the writing of this paper. The DDC staff involved with its Remote Access On-Line Information Retrieval System, particularly Mr. Richard Bennertz and Mr. Clinton LeMasters, were very generous with their time spent in amplifying the information contained in the DDC Remote Access System Manual. Mr. Van Wentes and Mrs. Madeline W. Losee of the NASA staff were particularly helpful in providing detailed information concerning NASA's RECON system. Without the effort of Mr. Kurt W. Stabenau in establishing the DDC Remote On-Line Terminal at NSRDC, the project would not have been possible.

Finally, the writer is greatly indebted to Mr. Abel W. Camara, Head, Information Retrieval Division, for his generous guidance, encouragement, and great patience.

REFERENCES

1. Defense Documentation Center, Draft of DDC Remote On-Line Retrieval System Manual, November, 1969, Washington, D. C., 1969.

2. Burko, H., "Interactive Document Storage And Retrieval System-Design Concepts", Mechanized Information Storage, Retrieval and Dissemination, Samuelson, K., ed., Amsterdam, North-Holland Publishing Co., 1968, pp. 591-599.

3. National Aeronautics and Space Administration, Introducing NASA's RECON, Washington, D. C., May, 1969.

4. Kesseler, M. M., TIP User's Manual, 2nd ed. rev., Cambridge, Mass., Massachusetts Institute of Technology, 1968.

5. Landau, R., et al, "On-Line Interactive Information Systems", Proceedings of the Sixth Annual National Colloquium on Information Retrieval, Schultz, L., ed., Philadelphia, Medical Documentation Service of the College of Physicians, 1969, pp. 359-372.

6. Salton, G., Automatic Information Organization and Retrieval, New York, McGraw-Hill, 1968.

7. Schneider, J. H., "Design of a Comprehensive Information System Based on Linear Hierarchical Decimal Classifications", Proceedings of the Sixth Annual National Colloquium on Information Retrieval, Schultz, L., ed., Philadelphia, Medical Documentation Service of the College of Physicians, 1969, pp. 99-105.

8. Driscoll, L., et al, "Inferential Retrieval", Proceedings of the Sixth Annual National Colloquium on Information Retrieval, Schultz, L., ed., Philadelphia, Medical Documentation Service of the College of Physicians, 1969, pp. 373-392.

9. Slagle, J., Textbook of Artificial Intelligence: The Heuristic Programming Approach, New York, McGraw-Hill, In Press.

10. Heaps, H. S. and Ko, W. C. C., "Automatic Adaptive Processing of Questions in Document Retrieval", Proceedings of the American Society for Information Science, Vol. 7: pp. 319-321, (1970).

11. Maron, M. E., "A Logician's View of Language-Data Processing", Natural Language and the Computer, Garvin, P. L., ed., New York, McGraw-Hill, 1963, pp. 128-150.

12. Hayes, R. M. "Mathematical Models in Information Retrieval", Natural Language and the Computer, Garvin, P. L., ed., New York, McGraw-Hill, 1963, pp. 268-309.